# SECOND-BEST CONGESTION PRICING SCHEMES IN THE MONOCENTRIC CITY

Erik T. Verhoef*
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
Email: everhoef@econ.vu.nl

This version: 08/05/04

**Key words:** Traffic congestion, second-best pricing, urban structure, spatial general equilibrium
**JEL codes:** R41, R48, D62

***Abstract***
*This paper considers second-best congestion pricing in the monocentric city, with endogenous residential density and endogenous labour supply. A spatial general equilibrium model is developed that allows consideration of the three-way interactions between urban density, traffic congestion and labour supply. Congestion pricing schemes are analyzed that are second-best 'by design' (and not because distortions exist elsewhere in the spatial economy), like cordon charging and flat kilometre charges. Both for Cobb-Douglas utility and for CES utility, the analyses suggest that the relative welfare losses from second-best pricing, compared to first-best pricing, are surprisingly small.*

*Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam.

## 1. Introduction

Practical applications of traffic congestion pricing typically (if not always) involve so-called second-best pricing regimes, which fail to charge every individual road user his or her exact marginal external congestion costs. With pay-lanes, to an increasing extent employed in the US, unpriced congestion remains existent on parallel highway lanes. In case of cordon charges, such as used in Singapore, every road user passing the cordon pays the same charge independent of the route followed before and after passing the cordon, and users who remain within or outside the cordon are exempted from paying the charge. Area charges, as recently introduced in London, impose the same charge on every user who drives within the area independent of the number of kilometres travelled, and leaves congestion outside the area uncharged. And flat kilometre charges, as currently considered for The Netherlands, do not differentiate by the place of driving and route followed.

A substantial literature has recently emerged on the economics of second-best congestion charges (*e.g.* Lindsey and Verhoef, 2001, provide an overview). Most of these studies employ partial equilibrium approaches, in which only the transport (network) market is considered explicitly. An exception is the work by authors such as Mayeres and Proost (2001) and Parry and Bento (2001), who study traffic congestion and road pricing for commuters in general equilibrium settings, allowing for distortions on the labour market. Their results suggest that these interactions can be of significant importance for the efficiency impacts of both congestion pricing and the use of the associated revenues.

Another non-transport market that is of importance when evaluating congestion pricing strategies for urban areas is the (spatial) housing market. Already in the 1970's, a number of studies appeared that looked into the interactions between traffic congestion and urban structure in the context of the monocentric model (Solow and Vickrey, 1971; Solow, 1972; Kanemoto, 1976; Arnott, 1979). Anas and Kim (1996) and Anas and Xu (1999) extended this line of research by allowing for multicentric configurations, endogenizing the emergence of centres through the explicit consideration of agglomeration forces.

The present paper aims to consider second-best congestion pricing in the monocentric city, with endogenous residential density and endogenous labour supply. A spatial general equilibrium model is developed that allows consideration of the three-way interactions between urban density, traffic congestion and labour supply. The model would therefore, for example, allow an investigation of second-best congestion pricing with distorted spatial labour markets. This matter, however, will be addressed in a companion paper to the present one (Verhoef, 2004). The present paper will instead be concerned with congestion pricing schemes that are second-best 'by design', like the examples mentioned above, and not because distortions exist elsewhere in the spatial economy.

Prior studies of traffic congestion in the monocentric model have typically looked at first-best congestion pricing measures, although second-best issues arising from non-optimal allocations of land to road capacity have been considered (*e.g.* Arnott, 1979). The recent contribution by Mun, Kunishi and Yoshikawa (2003) is an exception. They focus on second-

best optimal cordon pricing in a monocentric city. The policy appeared to perform unexpectedly well: when optimizing both the location of the cordon and the charge, welfare gains of around 94% of the gains from first-best pricing were found (computed from their Table 2). This is remarkably well when realizing that with a cordon charge, some road users will not face a congestion charge at all (those who live inside the cordon), some will face a charge that exceeds the marginal external costs they cause over their full trip (those who live outside but relatively close to the cordon), and a third group faces a charge below their marginal external costs (those who live outside and relatively far from the cordon).

Given the potentially far-reaching policy conclusions of this finding, an important question is to what extent the result depends on the assumed monocentric spatial configuration as such, and to what extent it is the result of other specific features of their model, such as the facts that urban densities are assumed given, and that a partial spatial equilibrium model was used. One might for instance hypothesize that an important difference between cordon charge and first-best tolls would be that the former provides a smaller marginal incentive to move closer to the city centre, as there is no reward in terms of a reduced congestion charge. At the same time, however, a cordon tax provides a relatively strong non-marginal, discrete, incentive to choose a location inside the cordon. The question arises whether, as a result of these opposing forces, the average density in the city increases or decreases under cordon charges compared to first-best tolling, and to which extent the discreteness of the charge and the likely resulting discontinuities of land rents and densities induce additional welfare losses due to cordon charging. One objective of this paper is to explore these questions by using a spatial general equilibrium model of a monocentric city. Compared to the model of Mun, Kunishi and Yoshikawa (2003), urban density will be made endogenous, trips will be assumed to involve commuting rather than other purposes (such as shopping), and only simultaneous equilibria of the transport market, the urban land market and the labour market will be considered. But as in Mun, Kunishi and Yoshikawa (2003), the monocentric urban structure is imposed exogenously, which means in the present model that all production is assumed to take place in a (spaceless) CBD. A model that endogenizes the formation of agglomerations within the urban area is planned for future work.

An alternative second-best policy is considered as well, and this involves flat kilometre charges. This means that also when marginal external congestion costs per kilometre driven vary over space, only a single per-kilometre congestion charge can be imposed throughout the city. Like cordon charges, such a policy could be motivated by excessive transaction and implementation costs for first-best congestion charges, typically requiring per-kilometre charges that vary continuously over space. Whereas the cordon tax does imply spatial variation of per-kilometre charges but at the cost of creating a discontinuity, the flat kilometre charge is in some sense its counterpart by avoiding discontinuities while preventing spatial variation of per-kilometre charges.

The two benchmarks against which both policies will be evaluated are the no-toll equilibrium on the one hand, and first-best congestion charging on the other.

## 2. The analytical model[1]

This section presents the details of the analytical model. Before turning to a detailed description of consumers' behaviour, the congestion technology, firms, and a characterization of general equilibrium, some introductory remarks are in order. First, $z$ will be used to denote a one-dimensional continuous urban space. The location of the spaceless CBD is at $z$=0, and the residential area stretches from $z$=0 to $z$=$z^*$, with $z^*$ being the endogenous city boundary. At the boundary of the city, the equilibrium residential bid-rent $r(z^*)$ should be equal to the exogenous and constant agricultural bid-rent $r_A$. A closed city is considered, meaning that the population size $N$ is treated as fixed and given.

It is assumed that all excess land rents above $r_A$ are redistributed among the city's population.[2] It is a convenient assumption in the sense that it easily allows us to consider households with similar initial endowments. As a result of these assumptions, some share of the urban production will not be consumed in the urban area, but will be exported in exchange for the purchase of land against the agricultural rent.

All consumers and producers are assumed to be price-takers. Households are identical, and so are firms. The industrial product can be transported costlessly, and the given world-market price of the industrial good $p$ is used as the *numéraire*. We now turn to the various actors in the city and the resulting equilibrium issues.

*Consumers*

The closed city has $N$ households, which are treated as a continuum of utility-maximizing economic entities. A household's utility depends on the consumption of the industrial good $y$, on the consumption of space or the size of the residence $s$, and on the consumption of free time or leisure $T_f$. A household's financial budget then consists of the net wage rate $w-\tau_L$ ($w$ is the gross wage, $\tau_L$ the labour tax) times the amount of hours worked $T_w$, plus the redistributed excess land rents ($R$ in total, $R/N$ per household), plus – possibly – a lump-sum government transfer $G$ (which is set so as to balance the government's budget). In equilibrium, the household's budget is fully spent on the consumption of $y$ and $s$, and – if levied – on road tolls and labour taxes. A household's given time budget is denoted by $T$, and can be spent on leisure ($T_f$), work ($T_w$) and commuting ($T_c$). All prices and taxes are treated parametrically by the (price-taking) households.

Commuting therefore does not require financial outlays other than possibly a total toll (*i.e.*, over the full trip) $\tau_R(z)$, but does take time $T_c(\cdot z)$ (the underlying travel time function will

[1] The exposition in this section closely follows that in Verhoef and Nijkamp (2002).

[2] Alternatively, an 'absentee land-lord assumption' could have been used, which assumes that none of the land rents generated in the city would be used for consumption in the city. Another possible assumption would be that *all* land rents generated in the city are redistributed among the population, which would in fact imply that the endogenous city size could – from the overall city's perspective – be expanded costlessly. The present representation compromises between these two polar cases, and would correspond to the situation where the public authority of the city buys the urban land against the relatively low rural land price, implying an equivalent (per-unit-of time) price of $r_A$, and redistributes all excess rents generated in the city among its population.

be discussed below). The number of commuting trips made by a household is assumed to be equal to the amount of effective working time supplied ($T_w$). Hence, $T_w$ is, as it were, expressed in terms of number of days worked.

A household's simultaneous labour supply and consumption decisions can be modelled by using the 'gross budget', that would be available under the maximum possible amount of time worked, and to let the household 'buy back' leisure time against the prevailing shadow price $w-\tau_L-\tau_R(z)$. Observing that the household's optimization problem is dependent on the residential location $z$, it can then be written as:

$$\begin{aligned} &\underset{y(z),s(z),T_f(z)}{Max}\ U\big(y(z),s(z),T_f(z)\big) \\ &s.t.\ \frac{R}{N}+G+\big(w-\tau_L-\tau_R(z)\big)\cdot\big(T-T_c(z)-T_f(z)\big)-p\cdot y(z)-r(z)\cdot s(z)=0 \end{aligned} \tag{1}$$

with:

$$R=\int_0^{z^*} r(z)-r_A\,\mathrm{d}z \tag{2a}$$

Both labour taxes and road tolls are collected and redistributed by the local government. A balanced government budget therefore implies that:

$$G=\frac{1}{N}\cdot\left(\int_0^{z^*} n(z)\cdot\big(T-T_c(z)-T_f(z)\big)\cdot\big(\tau_L+\tau_R(z)\big)\mathrm{d}z\right) \tag{2b}$$

where $n(z)$ gives the density of households at $z$. The 'gross budget' available at location z is thus defined as:

$$M(z)=\frac{R}{N}+G+\big(w-\tau_L-\tau_R(z)\big)\cdot\big(T-T_c(z)\big) \tag{3}$$

A spatial equilibrium requires that utility $U(z)$ be constant over $z$ for all $0<z\leq z^*$ (and exceeds $U(z)$ for $z>z^*$). This implies a particular equilibrium pattern of land-rents. We can be explicit about this when postulating a specific form for the utility function. Two types of utility function will be considered in this paper: Cobb-Douglas (with a unitary elasticity of substitution) and CES (constant elasticity of substitution). In this analytical section, only the Cobb-Douglas function is considered, which allows for an analytical expression for equilibrium land rents. It is expressed as:

$$\begin{aligned} &U(z)=y(z)^{\alpha_y}\cdot s(z)^{\alpha_s}\cdot T_f(z)^{\alpha_f} \\ &\text{with}:\quad \alpha_y+\alpha_s+\alpha_f=1 \end{aligned} \tag{4}$$

This utility function has the specific property of a unitary elasticity of substitution, implying that the gross budget shares spent on $y$, $s$ and $T_f$ will be constant, and given by the parameters $\alpha$. Specifically, the conditional demands for $y$, $s$ and $T_f$ are:

$$y(z) = \frac{\alpha_y \cdot M(z)}{p} \tag{5a}$$

$$s(z) = \frac{\alpha_s \cdot M(z)}{r(z)} \tag{5b}$$

$$T_f(z) = \frac{\alpha_f \cdot M(z)}{w - \tau_L - \tau_R(z)} \tag{5c}$$

and the indirect utility – for analytical convenience defined as the logarithm of the maximum utility achievable under given prices and wage – can be written as:

$$\begin{aligned} V(z) = {} & \alpha_y \cdot \ln \alpha_y + \alpha_s \cdot \ln \alpha_s + \alpha_f \cdot \ln \alpha_f + \ln\left(\frac{R}{N} + G + \left(w - \tau_L - \tau_R(z)\right) \cdot \left(T - T_c(z)\right)\right) \\ & - \alpha_y \cdot \ln p - \alpha_s \cdot \ln r(z) - \alpha_f \cdot \ln\left(w - \tau_L - \tau_R(z)\right) \end{aligned} \tag{6}$$

The condition that $V$ in (6) be constant over space implies:

$$V'(z) = \frac{-\tau_R'(z) \cdot \left(T - T_c(z)\right) - T_c'(z) \cdot \left(w - \tau_L - \tau_R(z)\right)}{\frac{R}{N} + G + \left(w - \tau_L - \tau_R(z)\right) \cdot \left(T - T_c(z)\right)} - \alpha_s \cdot \frac{r'(z)}{r(z)} + \alpha_f \cdot \frac{\tau_R'(z)}{w - \tau_L - \tau_R(z)} = 0 \tag{7}$$

where a prime denotes a 'space derivative' (with respect to location). Equation (7) gives a first-order differential equation for $r(z)$ that can be solved to yield:

$$r(z) = K \cdot \left(w - \tau_L - \tau_R(z)\right)^{-\frac{\alpha_f}{\alpha_s}} \cdot \left(R + G \cdot N + N \cdot \left(T - T_c(z)\right) \cdot \left(w - \tau_L - \tau_R(z)\right)\right)^{\frac{1}{\alpha_s}} \tag{8}$$

where K is a constant of integration. Invoking the equilibrium boundary condition that $r(z^*)=r_A$, we can solve for K:

$$K = \frac{r_A}{\left(w - \tau_L - \tau_R(z^*)\right)^{-\frac{\alpha_f}{\alpha_s}} \cdot \left(R + G \cdot N + N \cdot \left(T - T_c(z^*)\right) \cdot \left(w - \tau_L - \tau_R(z^*)\right)\right)^{\frac{1}{\alpha_s}}} \tag{9}$$

We conclude this part of the analysis with a few identities. We can find the local population density $n(z)$ as the inverse of the 'lot-size' $s(z)$:

$$n(z) = \frac{1}{s(z)} \tag{10}$$

The total population is given, so that:

$$\int_0^{z^*} n(z)\, dz = \int_0^{z^*} \frac{1}{s(z)}\, dz = N \tag{11}$$

Total labour supplied equals:

$$L = \int_0^{z^*} n(z) \cdot \left(T - T_c(z) - T_f(z)\right) dz \tag{12}$$

Total local consumption of the city's product equals:

$$Y = \int_0^{z^*} n(z) \cdot y(z)\, dz \tag{13}$$

The total amount of land consumed in the city must be equal to $z^*$, which is by definition true:

$$z^* = \int_0^{z^*} n(z) \cdot s(z)\, dz = \int_0^{z^*} 1\, dz = z^* \tag{14}$$

*Travel times*

A single radial road of a given, constant capacity is used jointly by all households when commuting. The per-unit-of-distance travel time $t(z)$ at each point along the road depends on the local density of commuters, defined by the cumulative labour supply between $z$ and $z^*$. A linear travel time function is used, defined by two constants $t_0$ (the free-flow travel time for one unit of distance) and $t_1$ (the function's slope):

$$t(z) = t_0 + t_1 \cdot \int_z^{z^*} n(\zeta) \cdot \left(T - T_c(\zeta) - T_f(\zeta)\right) \mathrm{d}\zeta \tag{15}$$

Total commuting time from $z$ to the CBD can then be written:

$$\begin{aligned} T_c(z) &= z \cdot t_0 + t_1 \cdot \int_0^z \int_\psi^{z^*} n(\zeta) \cdot \left(T - T_c(\zeta) - T_f(\zeta)\right) \mathrm{d}\zeta\, \mathrm{d}\psi \\ &= z \cdot t_0 + t_1 \cdot \int_0^{z^*} \mathrm{Min}\{\zeta, z\} \cdot n(\zeta) \cdot \left(T - T_c(\zeta) - T_f(\zeta)\right) \mathrm{d}\zeta \end{aligned} \tag{16}$$

The congestion externality is clearly reflected in (16): $T_c(z)$ depends on labour supply at every location in the city. The same therefore holds for the maximum utility attainable at $z$; compare equations (1) and, specifically for Cobb-Douglas utility, (6). Because labour supply at $z$ in turn depends on $T_c(z)$ (compare (5c) for Cobb-Douglas utility), the congestion externality induces complex direct and indirect spatial interactions throughout the city, in terms of both equilibrium utility obtained and in terms of labour supply decisions.

Absent economic distortions other than the traffic congestion externality, first-best road pricing involves spatially differentiated per-unit-of distance charges equal to per-unit-of-distance marginal external congestion costs. A problem equivalent to determining these is to find the marginal external congestion costs from supplying one additional unit of labour at every location $z$, and to derive the total (over the full trip) optimal road prices for trips as a function of trip origin $z$.

An additional unit of labour supplied at $z$ increases $T_c(\zeta)$ by $t_1 \cdot z$ for $\zeta \geq z$, and by $t_1 \cdot \zeta$ for $\zeta < z$. The associated marginal external costs, in monetary terms, can be found by expressing these increases in commuting travel times in equivalent monetary variations in the gross budget $M(\zeta)$. Equation (3) shows that the relevant shadow price of (leisure and work) time is $w-\tau_L-\tau_R(\zeta)$ (note that the other two consumer prices, $p$ and $r(\zeta)$, are not directly dependent on $T_c(\zeta)$). This shadow price is therefore in part directly dependent on the government's use of the two tax instruments. Because $n(\zeta)$ households will be affected at location $\zeta$, the marginal external costs of supplying one additional unit of labour at $z$, $mec(z)$ can therefore be written:

$$mec(z) = t_1 \cdot \int_0^{z^*} \mathrm{Min}\{\zeta, z\} \cdot n(\zeta) \cdot \left(w - \tau_L - \tau_R\right) \mathrm{d}\zeta \tag{17}$$

It is verified in the numerical model in Section 3 below that a policy of Pigouvian road taxes, defined by $\tau_R(z)=mec(z)$, with the revenues redistributed in a lump-sum manner via $G$, indeed maximizes equilibrium utility in the city when the labour $\tau_L$ is zero and hence no distortions in the labour market are present.

*Producers*

Probably the simplest possible production structure is assumed for the city. There is a continuum of firms, each of which is infinitesimally small relative to the market and takes all prices as given. The industrial output is homogeneous. All firms are located in the spaceless CBD, but the agglomeration forces that induce this clustering are not modelled explicitly. This also means that no market distortions through agglomeration externalities are assumed to exist; internal consistency could be achieved by, for example, assuming that zoning regulation prohibits firm location outside the CBD. The assumption of exogenous, central firm locations is clearly an unattractive feature of the present model, which is, however, made solely to allow us to concentrate on the performance of second-best congestion pricing in a monocentric city without introducing additional market distortions arising from agglomeration externalities. Because these market distortions are expected to be relevant in reality, they are considered explicitly in the companion paper Verhoef (2004).

Firms have a simple linear production technology with one input (labour). A firm's production function thus exhibits constant returns to scale, and therefore qualifies for application of Euler's theorem. The following aggregate production function applies:

$$Q = A \cdot L \tag{18}$$

Perfect competition drives profits to zero, with the result that the following equality holds:

$$p \cdot A = w \tag{19}$$

*General spatial equilibrium*

The model described above has 17 unkowns, some of which are functions of z. These unknowns are $V(z)$, $w$, $M(z)$, $r(z)$, $R$, $G$, $K$, $y(z)$, $Y$, $s(z)$, $n(z)$, $T_f(z)$, $L$, $z^*$, $t(z)$, $T_c(z)$, and $Q$ (recall that $r_A$, $p$ and $N$ are given; tax levels are treated exogenously and all other scalars are parameters). The 17 equations needed to solve this system are (2ab), (3), (5a-c), (6), (8)-(13), (15), (16), (18) and (19). For other types of utility and production functions, as long as they imply unique conditional (factor) demands, a similar equality of numbers of equations and unkowns should in principle hold. We refrain from a formal analysis of existence, uniqueness and stability of equilibria and optima in our model.

In our list of equations, we did not include the 'aggregate demand equals aggregate supply' relation, which in our partly open system reads:

$$p \cdot (Q - Y) = r_A \cdot z^* \tag{20}$$

Equation (20) states that the value of the city's production in excess of its local consumption should be just sufficient to pay for the purchase of land against the exogenous terms of trade $r_A/p$. The share of local production not exported is consumed locally. The reason for not including this equilibrium condition explicitly is that it will be automatically satisfied under the zero profit condition and exhaustion of consumers' financial budgets – as in fact dictated by Walras' Law. To see why, first observe that zero profits imply that:

$$p \cdot Q = w \cdot L \tag{21}$$

The exhaustion of consumers' total financial income, in combination with the balanced government budget, implies (in aggregate terms) that the sum of redistributed land rents and wage income should be equal to the sum of expenses on the local product and rents:

$$\begin{aligned} &R + w \cdot L = p \cdot Y + r_A \cdot z^* + R \\ &\Rightarrow w \cdot L = p \cdot Y + r_A \cdot z^* \end{aligned} \tag{22}$$

Substitution of (22) into (21) immediately yields (20).

It is not possible to obtain any further analytical (equilibrium) results for our model. We therefore now move to the results of a numerical illustration,[3] to study the comparative static properties of the free-market and some second-best (and first-best) equilibria.

[3] The numerical model was written in Mathematica 5.0, and finds spatial equilibria by using a repeated nested approach, in which the various markets of interest are successively brought into equilibrium while keeping other prices fixed, until convergence is reached (the convergence criterion used in the different loops was set at $1 \cdot 10^{-7}$ for relative changes in key variables between successive iterations within each loop). Provided reasonable starting values are used, this takes (for given policy instrument levels) less than 30 seconds on a modern lap-top computer. The flatness of the utility plots $U(z)$ in Figure 1 provides a graphical illustration of the fact that the model succeeds in producing a spatial equilibrium in which no incentives for relocation remains.

## 3. First-best and second-best road pricing: numerical results

### *3.1. Calibration*

The basic numerical model used deploys the Cobb-Douglas utility function also used above. Although the model is of course a rather strong abstraction from reality, some effort was put in calibrating the model so that the main endogenous results bear resemblance to what is observed in reality. We therefore start this section by briefly discussing the calibration of the model. It should be borne in mind that the calibration aimed to produce reasonable base-case equilibrium outcomes for a model that is in the first place rather abstract, and that in the second place is calibrated under the conceptually 'clean' but practically unrealistic assumption of zero (labour and transport) taxes.

We start with the normalizations used. As stated, the given world-market price of the good produced in the city is used as a *numéraire*, and $p$ is set equal to 1. Also the agricultural land rent is assumed to be given, and units of space are chosen such that also the agricultural rent $r_A$=1. Next, units of time are chosen such that the total time endowment $T$=1. And finally, the number of households is set at 1000.

The parameters $\alpha$ of the Cobb-Douglas utility function determine the equilibrium (gross) budget shares of the industrial good, housing and leisure. The values were set at $\alpha_y$=0.2, $\alpha_s$=0.15 and $\alpha_f$=0.65. Because the monetary budget is fully spent on the industrial good and on housing, the first two $\alpha$'s imply that some 43% of total monetary income is spent on housing and 57% on other consumption. This seems reasonable for urban areas. The value of $\alpha_f$, in combination with equilibrium city size and commuting times, leads to an average $T_w$ of 0.29 (see also Figure 1 and Table 1 below). For an average week, consisting of 7·16 hours (excluding 8 hours sleep per day), this means 32 hours working time.

The final two parameters to be chosen are $t_0$, set at $7.5\cdot10^{-4}$, and $t_1$, set at $7.5\cdot10^{-6}$. The ratio of these two parameters causes the equilibrium speed near the CBD to be just over 25% of the free-flow speed (as applying at the city fringe). The absolute sizes of these parameters – together with the further parametrization – cause the equilibrium city size $z^*$ to be such that the person living at $z^*$ has a commuting time $T_c(z^*)$ of 0.57· $T_w(z^*)$. With labour supplied in units of 8-hour working days, this would mean a maximum commute of some 4.5 hour per working day (for a return trip), or 2.25 hour for a single trip.

Finally, in the base equilibrium, wage income forms some 84.5% of total monetary income (the remainder being redistributed excess rents). Figure 1 show the equilibrium spatial patterns of some further variables of interest. Note for example that residential land rent near the CBD is around three times as high as the agricultural rent and lot sizes are around 2.5 times as small. The convex shape of the equilibrium land rent is caused by both the possibility of substitution in consumption, and the fact that equilibrium per-unit-of distance transport costs increase towards the CBD.

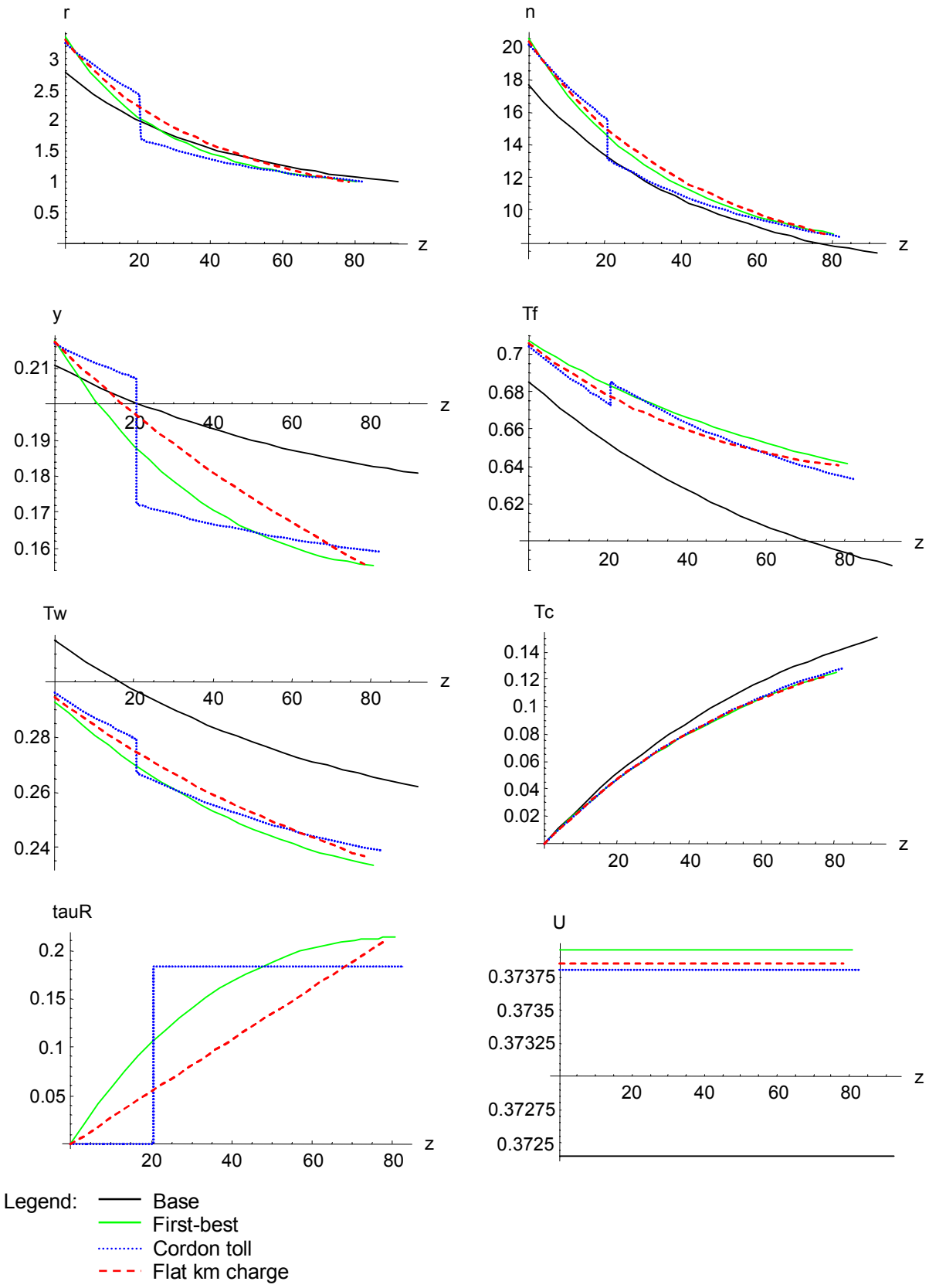


*Figure 1. Key results for Cobb-Douglas utility*

### *3.2. First-best and second-best road pricing: Cobb-Douglas utility*

The model is used to evaluate the performance of two second-best road pricing schemes, relative to the base equilibrium and the benchmark of optimal Pigouvian congestion charging. For each of the three road pricing schemes, it is assumed that toll revenues are redistributed as a lump-sum benefit $G$, which is the optimal type of redistribution given that no initial labour taxes are existent (the alternative of recycling through negative labour taxes would distort the labour-leisure trade-off). Table 1 summarizes the equilibrium levels of the model's main endogenous non-spatial variables, while Figure 1 compares spatial patterns for the main spatially differentiated variables. Before turning to these, it is useful to discuss the procedures used to find the optimal levels for the tax instruments.

*Finding first-best and second-best tax levels*

The first-best equilibrium was found by consistently applying space-varying per-unit-of-distance Pigouvian congestion taxes as given in (17) in the transport market:

$$\tau_R(z) = mec(z) \tag{23a}$$

The optimality of this policy (when revenues are redistributed in a lump-sum manner) was verified by investigating equilibrium utility levels for four perturbations of (23a). The first two applied tax rates $\tau_R(z)=0.9\cdot mec(z)$ and $\tau_R(z)=1.1\cdot mec$. The other two tested 'tilted' tax schedules according to $\tau_R(z)=(0.9+0.2\cdot z/z^*)\cdot mec$ and $\tau_R(z)=(1.1-0.2\cdot z/z^*)\cdot mec$. All perturbations led to utility levels below the first-level ($U$=0.373958), with deviations occurring only from the fifth digit onwards (the fifth and sixth digits became 48, 49, 44, and 43, respectively (in order of appearance of the perturbations above); compared to 58 for first-best pricing).

The second-best cordon tax requires the optimal choice of two instruments: the location of the cordon ($z_{cor}$) and the toll level ($\tau_{cor}$). The implied road toll becomes:

$$\tau_R(z) = \begin{cases} \tau_{cor} & \text{if } z \geq z_{cor} \\ 0 & \text{otherwise} \end{cases} \tag{23b}$$

As was the case for the model in Mun, Kunishi and Yoshikawa (2003), no closed-form solutions for the second-best optimal levels of $z_{cor}$ and $\tau_{cor}$ could be found, and a heuristic grid search method was used to identify these. The method entails two stages. First, 4·4 combinations of $z_{cor}$ and $\tau_{cor}$ were tested, and a 'utility hill' as shown in Figure 2 was constructed from the results by means of third-order interpolation. Its maximum entails the first-round prediction of $z_{cor}$ and $\tau_{cor}$. Next, the same procedure was applied to again 4·4 combinations of $z_{cor}$ and $\tau_{cor}$, where for both the range was chosen to be between –20% and +20% of the first-round predictions. The prediction of optimal values in this second round was taken to give the second-best optimal instrument levels. Due to the flatness of the 'utility hill' near the second-best optimum, further refinement would seem redundant.

Figure 2 shows that equilibrium utility appears to be relatively speaking more sensitive to deviations in $\tau_{cor}$ than to deviations in $z_{cor}$ (a similar pattern was also found for the CES utility function). This is in some sense good news for the design of cordon toll schemes when

the regulator is uncertain about the second-best optimal levels of $\tau_{cor}$ and $z_{cor}$. Whereas the location of the cordon will often involve relatively large fixed costs due to installation of necessary equipment, toll levels are in principle more flexible. The pattern shown in Figure 2 suggests that a relatively small mistake in the location of the cordon need not cause large relative welfare losses. The instrument for which mistakes are relatively speaking more important – the toll level – is also the one that is probably less costly to adjust in reality.

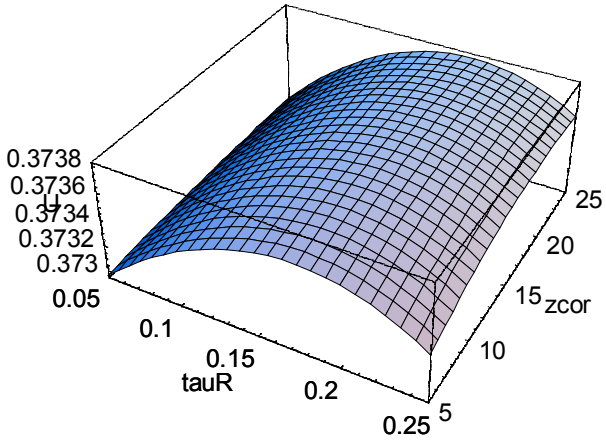


*Figure 2. 'Utility hill' for cordon charging*

The third congestion tolling scheme considered involves 'flat' (*i.e.*, not differentiated over space) kilometre charges. The tax rate, $\tau_{km}$ was again found by a heuristic procedure; a single-nested one-dimensional variant of the procedure used for cordon charges. The implied road tolls now simply becomes:

$$\tau_R(z) = z \cdot \tau_{km} \tag{23c}$$

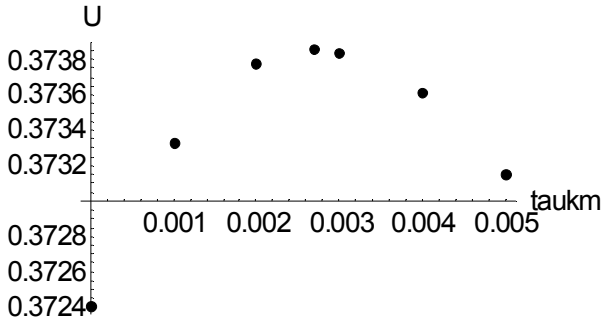


*Figure 3. 'Utility hill' for flat kilometre charges*

Figure 3 depicts the 'utility hill' for this policy, with the central point being the second-best optimum. Again, flatness of the utility hill near the second-best optimum secures that small deviations from the truly second-best optimal tax level will not affect the results in any significant way.

*The impacts and relative performance of first-best and second-best congestion pricing*

Figure 1 and Table 1 present the main findings. It is instructive to first consider the major impacts of first-best charging. This policy leads to reductions in aggregate kilometrage (*KM*) of around 21% and in aggregate commuting time (*TC*) of nearly 18%. These reductions are achieved partly by a 9% decrease in labour supply (and production), and a 12.4% decrease in city size (the latter explains why *KM* decreases more strongly than *TC*). The result is that utility increases by 0.42%; *i.e.*, introduction of optimal congestion pricing raises utility by the same amount as equiproportional increases of $y$, $s$ and $T_f$ by 0.42% would. This may seem a modest increase at first sight, but it appears to be reasonably in line with recent predictions of surplus gains from optimal road pricing in urban areas.[4]

| | Base equilibrium | First-best charging[a] | Cordon charging[a] | Flat km charging[a] |
|---|---|---|---|---|
| Policy instruments | | | | |
| $\tau_L$ | 0 | 0 | 0 | 0 |
| $\tau_R$ | 0 | see Fig. 1 | | |
| $\tau_{cor}$ | | | 0.183 | |
| $z_{cor}$ | | | 20.59 | |
| $\tau_{km}$ | | | | 0.00269 |
| $G$ | 0 | 0.03326 | 0.02939 | 0.02296 |
| Endogenous variables | | | | |
| $L$ | 286.866 | 90.95% | 92.29% | 92.71% |
| $Q$ | 286.866 | 90.95% | 92.29% | 92.71% |
| $Y$ | 194.847 | 92.52% | 93.44% | 96.31% |
| $z^*$ | 92.02 | 87.62% | 89.87% | 85.09% |
| $R$ | 54.115 | 100.84% | 99.50% | 115.40% |
| $KM$[b] | 10860.4 | 78.95% | 81.38% | 78.63% |
| $TC$[c] | 79.8824 | 82.16% | 83.72% | 81.38% |
| $TF$[d] | 633.251 | 106.35% | 105.56% | 105.65% |
| Toll revenues | 0 | 33.26 | 29.34 | 22.96 |
| Tax revenues | 0 | 0 | 0 | 0 |
| $U$ | 0.3724 | 100.42% | 100.38% | 100.39% |
| $\omega$ | 0 | 1 | 0.901 | 0.934 |

Notes:
[a] Percentages are relative to base equilibrium levels
[b] Aggregate kilometrage, defined as $\int z \cdot n(z) \cdot T_w(z)\, dz$
[c] Aggregate commuting time, defined as $\int n(z) \cdot T_c(z)\, dz$
[d] Aggregate leisure time, defined as $\int n(z) \cdot T_f(z)\, dz$

*Table 1. The relative impacts of first-best and second-best congestion pricing schemes*

Figure 1 confirms that residential density increases throughout the city, while central rent rise and rents near the fringe fall (the latter is consistent with the fringe rent for the smaller optimal city still being equal to the agricultural rent). The consumption of $y$ falls at nearly all locations (although close to the CBD the redistributed revenues dominate the toll payments

[4] Lindsey *et al.* (2004) provide estimates of annual per capita social surplus gains from first-best road pricing for four European cities (Paris, Brussels, Helsinki and Oslo), which vary from € 111 – 403. The average of around € 257 corresponds, in terms of our model, with 0.42% gain in a household's gross budget if it amounts to around € 61 000. If, as in our model, 35% of this gross budget is monetary, the households monetary income should be around € 21 500 to make the welfare gains from the present study comparable to those in Lindsey *et al.* (2004), which appears a reasonable order of magnitude.

and consumption increases), while the consumption of leisure increases as both labour supply and travel times fall. The concave spatial pattern of $\tau_R(z)$ confirms the intuitive notion that per-unit-of-distance tolls are zero at the city fringe and rise more than proportionally towards the CBD. Equilibrium utility of course remains constant over space, albeit that the level increases.

Both second-best policies perform rather well in terms of relative welfare gains. Both accomplish more than 90% of first-best gains, as shown by the efficiency indicator $\omega$ (defined as the proportion of the first-best equilibrium utility increase that the policy achieves).

The relative performance of the cordon charge, with $\omega$=0.901, is well in line with the findings of Mun, Kunishi and Yoshikawa (2003), who found $\omega$=0.940 for their basic model. Although the relative welfare losses have, of course, nearly doubled, neither the sharp kinks that the policy produces at $z_{cor}$ (see Figure 1), nor the inherent distortions resulting from the inability of charging all road users and from applying imperfect charges to (nearly) all others, appear to seriously undermine the welfare gains – even when accounting for induced changes in residential densities and labour supply decisions.

What causes cordon charges to be so efficient? Part of the explanation lies in the fact that it aggregate terms, the tax induces both a reduction in labour supply and a reduction in city size, which were the main two changes also induced by first-best taxes. The aggregate density increases (compared to the base equilibrium) for two reasons. First, land inside the cordon becomes relatively attractive, which drives up land rents and hence density (see Figure 1). Secondly, the toll discourages labour supply outside the cordon, which translates – via reduced budgets – into lower land consumption and hence a higher density. Labour supply decreases for two reasons. Inside the cordon, the redistributed toll revenues increase the gross budget, which encourages the consumption of leisure. Outside the cordon, the toll discourages labour supply. Therefore, the cordon tax does induce the same two aggregate responses as first-best tolls do. Of course, the kinks introduced by the cordon toll – illustrated in Figure 1 – lead to welfare losses compared to first-best prices. But the diagrams show that, provided the cordon location and toll level are set optimally, spatial patterns of key variables under cordon charging are nevertheless relatively close to first-best results, with 'too large' levels inside the cordon compensated by 'too low' levels outside the cordon, and *vice versa*.

The flat kilometre charge, with $\omega$=0.934, performs even better than the cordon charge. Again, the policy succeeds in reducing both labour supply and city size. The modesty of welfare losses compared to first-best pricing is now intuitively explained that the per-unit-of-distance tax rates are too low near the CBD, and too high near the fringe. Because households only consider the full-trip toll, there is a natural tendency for the two errors to cancel. Certainly, they can not cancel exactly for all $z$ (which is why $\omega$<1), but the optimal flat rate does a good job at finding a reasonably efficient compromise. In aggregate terms (Table 1), most results for flat km charges are comparable to those of cordon charging. The main exceptions are that optimal toll revenues are significantly lower, and aggregate excess land-rents are significantly higher. The latter also exceed first-best rents, which is due to the fact

that the equilibrium bid-rent is, as expected, less convex under flat charges than under first-best pricing, while the city size and the central rent are nearly identical.

All in all, when comparing the two second-best instruments, the efficiency losses from the inability to differentiate tolls over space (under flat km charges) are apparently somewhat lower than the efficiency gains from avoiding kinks. When comparing both to first-best pricing, it is especially the high relative efficiency that remains surprising. In combination with the results of Mun, Kunishi and Yoshikawa (2003), this raises the hypothesis that the regularity of the spatial lay-out of the monocentric city probably yields good opportunities for minimizing the inherent distortions from which second-best congestion charging mechanisms suffer. This raises the question of whether the monocentric model is the appropriate model for studying the relative performance of such policies in reality. If anything, a generalization of these favourable results to polycentric cities seems premature.

### *3.3. First-best and second-best road pricing: CES utility*

The Cobb-Douglas utility function deployed up to this point may be critized for its restrictive assumption of unitary elasticity of substitution. It is therefore of some interest to investigate the impacts of the same policies when a more general constant elasticity of substitution (CES) utility function applies. This leads to changes in equations (4)–(9) above. Using primes to denote the relevant 'CES' equations, the relevant formulations become:

$$U(z) = \left( \left(\delta_y \cdot y(z)\right)^{\rho} + \left(\delta_s \cdot s(z)\right)^{\rho} + \left(\delta_f \cdot T_f(z)\right)^{\rho} \right)^{\frac{1}{\rho}} \tag{4'}$$

The elasticity of substitution, $\sigma$, is equal to $1/(1-\rho)$; while a convenient parameter when working with this type of utility function is $\chi = \rho/(\rho-1)$. The conditional demands for $y$, $s$ and $T_f$ become:

$$y(z) = M(z) \cdot \frac{\left( \frac{p}{\delta_y{}^{\rho}} \right)^{-\sigma}}{\left( \frac{p}{\delta_y} \right)^{\chi} + \left( \frac{r(z)}{\delta_s} \right)^{\chi} + \left( \frac{w - \tau_L - \tau_R(z)}{\delta_f} \right)^{\chi}} \tag{5a'}$$

$$s(z) = M(z) \cdot \frac{\left( \frac{r(z)}{\delta_s{}^{\rho}} \right)^{-\sigma}}{\left( \frac{p}{\delta_y} \right)^{\chi} + \left( \frac{r(z)}{\delta_s} \right)^{\chi} + \left( \frac{w - \tau_L - \tau_R(z)}{\delta_f} \right)^{\chi}} \tag{5b'}$$

$$T_f(z) = M(z) \cdot \frac{\left( \frac{w - \tau_L - \tau_R(z)}{\delta_f{}^{\rho}} \right)^{-\sigma}}{\left( \frac{p}{\delta_y} \right)^{\chi} + \left( \frac{r(z)}{\delta_s} \right)^{\chi} + \left( \frac{w - \tau_L - \tau_R(z)}{\delta_f} \right)^{\chi}} \qquad (5c')$$

The indirect utility can be written as (while writing $M(z)$ in full):

$$V(z) = \left( \frac{R}{N} + G + \left(w - \tau_L - \tau_R(z)\right) \cdot \left(T - T_c(z)\right) \right) \cdot \left( \left( \frac{p}{\delta_y} \right)^{\chi} + \left( \frac{r(z)}{\delta_s} \right)^{\chi} + \left( \frac{w - \tau_L - \tau_R(z)}{\delta_f} \right)^{\chi} \right)^{\frac{-1}{\chi}} \qquad (6')$$

The space-derivative of $V$ in (6′) is a straightforward but tedious expression and is therefore suppressed, while for the implied first-order differential equation for $r(z)$ – to obtain a constant utility over space – no closed-form analytical solution could be obtained. Numerical solutions, however, could be obtained, and these will be reported below.

*Calibration*

To obtain sufficient contrast with the Cobb-Douglas utility function, the elasticity of substitution $\sigma$ was set equal to 0.5 (with the corresponding $\rho$ and $\chi$ following as indicated in (4′)). To maximize comparability with the Cobb-Douglas model, all non-utility parameters were kept unchanged. Only the parameters $\delta$ therefore had to be calibrated, and these were set such that with weighted average prices from the Cobb-Douglas equilibrium ($\bar{p} = 1$, $\bar{r} = (R + z^* \cdot r_A)/z^*$ and a shadow price of leisure $\bar{p}_f = w - \tau_L -$ total toll revenues over total leisure), the budget shares from the CES utility function are equal to those from the Cobb-Douglas function. Some basic manipulations reveal that this is achieved when, for consumption good $x$ with a weighted average price $\bar{p}_x$, $\delta_x$ from the CES function is related to $\alpha_x$ from the Cobb-Douglas function according to $\delta_x = \bar{p}_x / \alpha_x{}^{\frac{1}{\chi}}$. This yielded $\delta_y$=25.0; $\delta_s$=70.6 and $\delta_f$=2.37. The base-case results in Table 2 confirm that the aggregate levels of equilibrium variables with the CES utility function are indeed close to those with the Cobb-Douglas function in Table 1. The spatial patterns shown in Figure 4 also reflect a close correspondence with the Cobb-Douglas results, albeit that the lower elasticity of substitution now causes a less pronounced differentiation of quantities consumed over space.

*First-best and second-best congestion pricing*

Table 2 shows the results of first-best and second-best congestion pricing for the CES utility function. The main conclusions are (1) that relative gains from congestion pricing, compared to the base equilibrium, are lower than under Cobb-Douglas utility due to the reduced sensitivity of households to price differences; and (2) the relative welfare gains of the two second-best policies, compared to first-best welfare gains, are nearly identical to those under Cobb-Douglas utility. In other words, the lower elasticity of substitution affects the size of

welfare gains from congestion pricing, but not the relative welfare gains from different pricing schemes.

| | Base equilibrium | First-best charging[a] | Cordon charging[a] | Flat km charging[a] |
|---|---|---|---|---|
| Policy instruments | | | | |
| $\tau_L$ | 0 | 0 | 0 | 0 |
| $\tau_R$ | 0 | see Fig. 2 | | |
| $\tau_{cor}$ | | | 0.208 | |
| $z_{cor}$ | | | 24.03 | |
| $\tau_{km}$ | | | | 0.00282 |
| $G$ | 0 | 0.03920 | 0.03475 | 0.02774 |
| Endogenous variables | | | | |
| $L$ | 284.552 | 94.69% | 95.37% | 95.66% |
| $Q$ | 284.552 | 94.69% | 95.37% | 95.66% |
| $Y$ | 193.509 | 95.48% | 95.88% | 97.60% |
| $z^*$ | 91.0437 | 93.01% | 94.29% | 91.53% |
| $R$ | 57.9265 | 108.58% | 107.19% | 125.91% |
| $KM^b$ | 11228.9 | 87.73% | 89.12% | 87.58% |
| $TC^c$ | 86.5438 | 89.76% | 90.73% | 89.31% |
| $TF^d$ | 628.904 | 103.81% | 103.37% | 103.43% |
| Toll revenues | 0 | 39.20 | 34.76 | 27.74 |
| Tax revenues | 0 | 0.00 | 0.00 | 0.00 |
| $U$ | 0.964227 | 100.26% | 100.24% | 100.24% |
| $\omega$ | 0 | 1 | 0.906 | 0.934 |

Notes:
[a] Percentages are relative to base equilibrium levels
[b] Aggregate kilometrage, defined as $\int z \cdot n(z) \cdot T_w(z)\, dz$
[c] Aggregate commuting time, defined as $\int n(z) \cdot T_c(z)\, dz$
[d] Aggregate leisure time, defined as $\int n(z) \cdot T_f(z)\, dz$

*Table 2. The relative impacts of first-best and second-best congestion pricing schemes: CES utility*

## 4. Conclusion

The results presented in this paper suggest that the surprisingly optimistic conclusions that Mun, Konishi and Yoshikawa (2003) reach on the relative performance of cordon pricing in the monocentric city are robust with respect to the inclusion of residential land markets, endogenous labour supply and general spatial equilibrium formulation conditions. Moreover, the result is obtained both for Cobb-Douglas and CES utility functions. This raises the suspicion that the regular monocentric configuration, which moreover ignores transport network effects, may be the responsible factor for this counter-intuitive result. The analysis furthermore showed that also flat kilometre charges perform surprisingly well in this setting – even slightly outperforming cordon charges.

A future research agenda is easily sketched. One line of research would endogenize the formation of (sub-)centres, by endogenizing firm location decisions and agglomeration advantages. A second line of research enabled by the model proposed in this paper concerns the investigation of second-best distortions in congestion pricing as arising from the existence of distortionary labour taxes in a spatial general equilibrium setting.

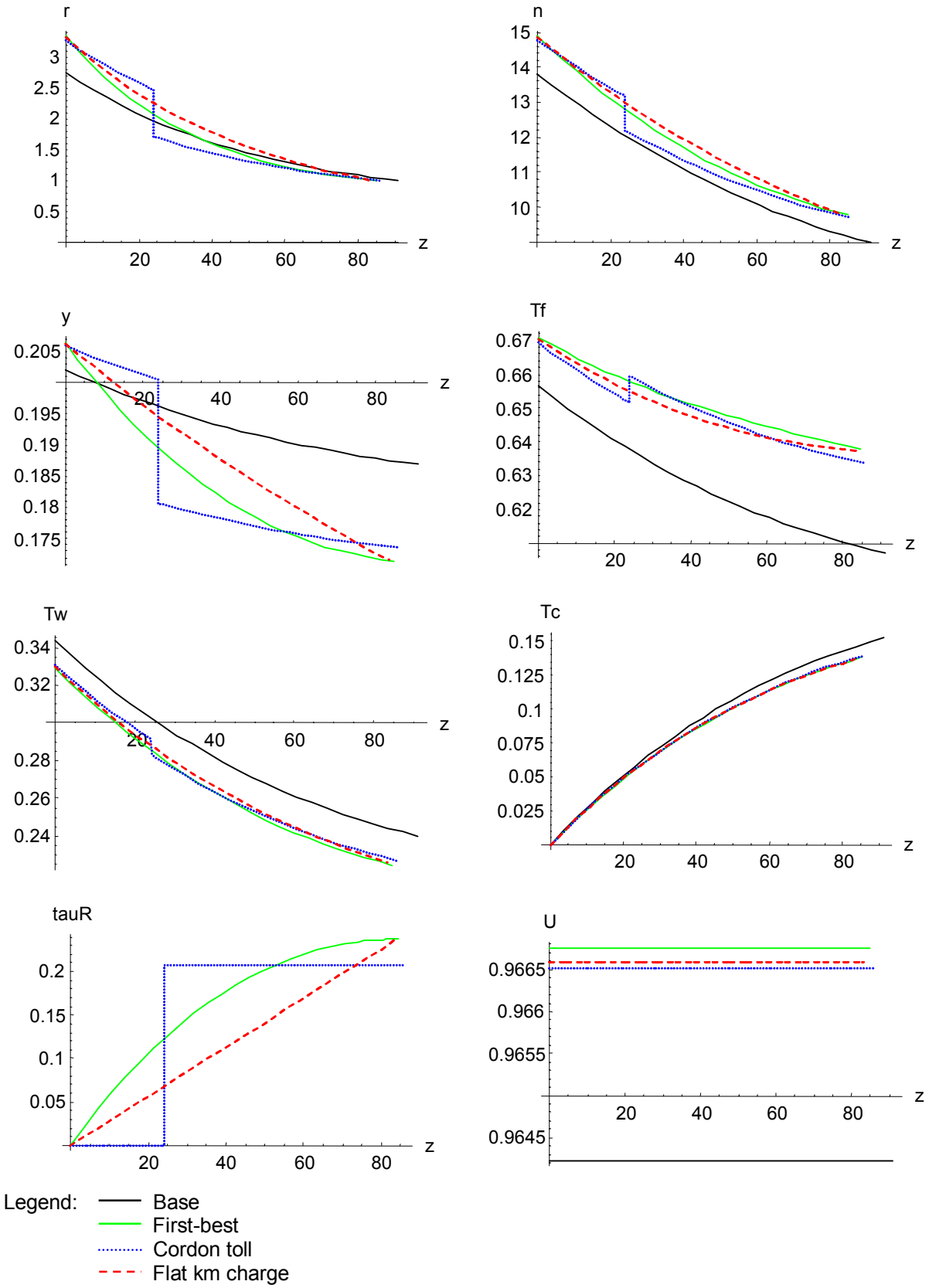


*Figure 4. Key results for CES utility*

## References

Anas, A., and I. Kim (1996) "General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration" *Journal of Urban Economics* **40** 232-256.

Anas, A., and R. Xu (1999) "Congestion, land use and and job dispersion: a general equilibrium model" *Journal of Urban Economics* **45** 451-473.

Arnott, R. (1979) "Unpriced transport congestion" *Journal of Economic Theory* **21** 294-316.

Kanemoto, Y. (1976) "Cost-benefit analysis and the second-best land use for transportation" *Journal of Urban Economics* **4** 483-503.

Lindsey, C.R. and E.T. Verhoef (2001) "Traffic congestion and congestion pricing". In: D.A. Hensher and K.J. Button (eds.) (2000) *Handbook of Transport Systems and Traffic Control, Handbooks in Transport 3* Elsevier / Pergamon, Amsterdam, pp. 77-105.

Lindsey, C.R., E. Niskanen, E.T. Verhoef, A. de Palma, P. Moilanen, S. Proost and A. Vold (2004) "Implementation paths for marginal-cost-based pricing in urban transport: theoretical considerations and case study results" unpublished manuscript.

Mayeres, I. and S. Proost (2001) "Marginal tax reform, externalities and income distribution" *Journal of Public Economics* **79** 343-363.

Mun, S., K. Konsihi and K. Yoshikawa (2003) "Optimal cordon pricing" *Journal of Urban Economics* **54** 21-38.

Parry, I.W.H. and A.M. Bento (2001) "Revenue recycling and the welfare effects of congestion pricing" *Scandinavian Journal of Economics* **103** 645-671.

Solow, R.M. (1972) "Congestion, density and the use of land in transportation" *Swedish Journal of Economics* **74** 161-173.

Solow, R.M. and W. Vickrey (1971) "Land use in a long, narrow city" *Journal of Economic Theory* **3** 430-447.

Verhoef (2004) "Traffic congestion and spatial labour markets" Dept. of Spatial Economics, Free University Amsterdam (currently in progress).

Verhoef, E.T. and P. Nijkamp (2002) "Externalities in urban sustainability: environmental versus localization-type agglomeration externalities in a general spatial equilibrium model of a single-sector monocentric industrial city" *Ecological Economics* **40** 157-179.